SciDataCon 2016

Building a Strategy for Data Management and Preservation to Support Agricultural Research

Marcos Cezar Visoli, Debora Pignatari Drucker, Patricia Rocha Bella Bertin, Embrapa

Summary

Digital scientific data arising from heterogeneous sources and techniques, such as ground measurements and remotely-sensed data, bring together many challenges associated with its curation, preservation, integration and dissemination. Within large, geographically-distributed organizations, such as the Brazilian Agricultural Research Corporation (Embrapa), avoiding redundant efforts on data acquisition and management becomes a difficult task, as different research centers and communities develop equally diverse data management tools and culture. Frequently, data practices do not conform to well-stablished, widely-used standards, hampering data integration and reuse. This paper describes Embrapa's strategy to promote best practices on documentation, preservation, reuse and integration of data throughout its lifecycle, as a means to foster agricultural knowledge production and sharing. The corporate strategy was developed in a participatory and collaborative way, taking into account the multiple perspectives of all relevant actors. For this purpose, an internal survey was conducted that allowed a better understanding of the main data management practices and cultures across the organization. Survey results served as a basis to the formulation of a corporate strategy and recommendation of best management practices throughout the research data lifecycle, which accorded to five main components of digital data infrastructures: people, technology, data, institutional framework and standards. The overall process employed for building Embrapa's research data management strategy provides a useful analytical framework that can be easily adapted for other organizations and contexts.

Keywords: Agriculture, Research Data, Data Life Cycle, Data Preservation, Data Reuse, Data Quality

The challenge of managing research data in a large, geographically distributed organization

Embrapa, the Brazilian Agricultural Research Corporation, is a governmental organization composed by 46 research centers, which are widely distributed across the country (Figure 1). The purpose of the organization is to promote Research, Development and Innovation for sustainable agriculture.

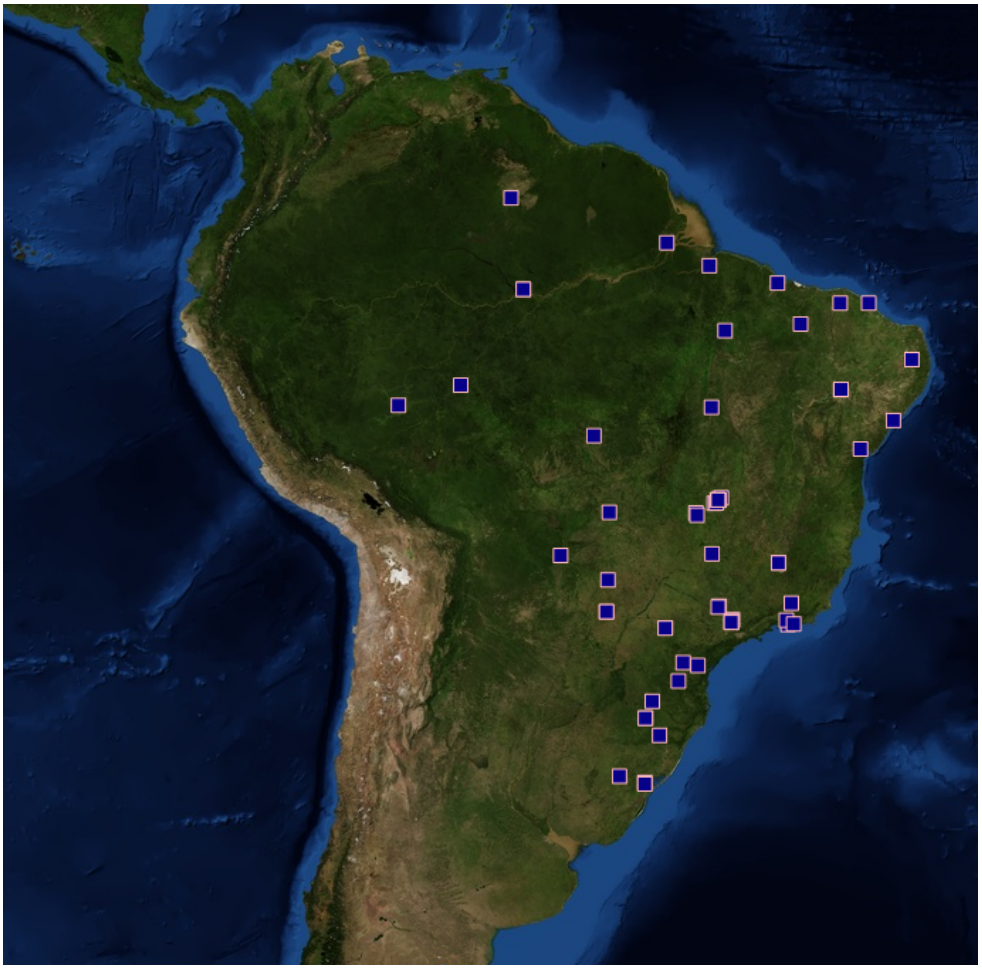

Figure 1: Geographic distribution of the 46 research centers of the Brazilian Agricultural Research Corporation (Embrapa), a governmental organization. Research centers are represented by blue squares.

Developing a corporate data management program to suit such a large and complex organization, with scientific data emerging from heterogeneous sources and techniques, and in accordance with a variety of agricultural research specialisms and cultures present many challenges. As distributed research centers enable broad access to experimental fields and sites across the country, different research communities tend to develop equally diverse data management tools and culture over time.

The overall approach

In order to face such challenges, it was first necessary to achieve a better understanding of current research data management practices and initiatives across the organization. Since the beginning, it was defined that Embrapa's strategy for research data management should be developed in a participatory and collaborative way, taking into account the multiple perspectives of all relevant actors. To achieve so, an internal survey was conducted, during six regional workshops, which allowed identifying the gaps and best practices across different research groups and centers.

Two representatives of each one of Embrapa's research centers took part in the regional workshops. After discussing the main concepts and principles of research data management,

respondents filled in an electronic questionnaire covering research data practices throughout its lifecycle (Strasser et al. 2011): collect, assure, describe, discover, preserve, integrate, analyze.

Preliminary results and way forward

The survey produced detailed information about the diverse data managing cultures and strategies adopted by different research centers and groups at Embrapa. In some cases, practices adopted do not meet international, widely-used standards, hampering data integration and reuse. It was found, for instance, that while some continued experiments with long research tradition usually count on relational databases tailored to specific data usage needs, data collections strongly based on human observation are frequently managed in non-structured spreadsheets.

Results not only allowed the identification of best practices and gaps, but also served as a basis for the recommendation of actions in the short and long term that will improve research data management throughout the organization. The proposed actions were organized as a roadmap and categorized according to the main components of digital data infrastructures: people, technology, data, institutional framework and standards. Among the several recommendations is the institutionalization of well-established and widely-adopted standards to support data interoperability, as well as the adoption of international principles for research data management such as those promoted by the DAMA-DMBOK (DAMA International 2010).

Besides providing material upon which a roadmap for an institutional research data management program was designed, the web survey and the regional workshops contributed for raising awareness on the subject, while providing an opportunity to discuss the main issues, principles and foundations of research data management.

Tackling the complex challenges with which agricultural science is faced nowadays requires accessing trustable data sources. Although still in its infancy, Embrapa's experience in building a corporate research data management program provides a useful analytical framework that can be easily adapted for other organizations and contexts.

Acknowledgements

The authors would like to thank all the survey participants for their contribution as survey respondents and for their insightful perceptions and suggestions to promote data management best practices throughout the data lifecycle. The authors thank Embrapa directory board for their support on building the strategy described at this paper.

Competing Interests

The authors declare that they have no competing interests.

References

DAMA International. 2010. *The DAMA Guide to the Data Management Body of Knowledge (DAMA-DMBOK Guide).* 1st. Technics Publications.

Strasser, C.; Cook, R.; Michener, W.; Budden, A.; Koskela, R. 2011. DataONE: promoting data stewardship through best practices. In: PROCEEDINGS OF THE ENVIRONMENTAL INFORMATION MANAGEMENT CONFERENCE, 2011. California: University of California. p. 126–131.